The Goal

• Find “patterns”: local regularities that occur more often than you would expect. Examples:
  • If a person buys wine at a supermarket, they also buy cheese. (confidence: 20%)
  • If a person likes Lord of the Rings and Star Wars, they like Star Trek (confidence: 90%)
• Look like they could be used for classification, but
• There is not a single class label in mind. They can predict any attribute or a set of attributes. They are unsupervised
• Not intended to be used together as a set
• Often mined from very large data sets
Example Data

Market basket analysis, e.g., supermarket

<table>
<thead>
<tr>
<th>Chicken</th>
<th>Onion</th>
<th>Rocket</th>
<th>Caviar</th>
<th>Haggis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

These are databases that companies have already.

Transactions trip to market

Thursday, 8 March 12
Other Examples

- Collaborative-filtering type data: e.g., Films a person has watched
- Rows: patients, columns: medical tests (Cabena et al, 1998)
- Survey data (Impact Resources, Inc., Columbus OH, 1987)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Demographic</th>
<th># Values</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sex</td>
<td>2</td>
<td>Categorical</td>
</tr>
<tr>
<td>2</td>
<td>Marital status</td>
<td>5</td>
<td>Categorical</td>
</tr>
<tr>
<td>3</td>
<td>Age</td>
<td>7</td>
<td>Ordinal</td>
</tr>
<tr>
<td>4</td>
<td>Education</td>
<td>6</td>
<td>Ordinal</td>
</tr>
<tr>
<td>5</td>
<td>Occupation</td>
<td>9</td>
<td>Categorical</td>
</tr>
<tr>
<td>6</td>
<td>Income</td>
<td>9</td>
<td>Ordinal</td>
</tr>
<tr>
<td>7</td>
<td>Years in Bay Area</td>
<td>5</td>
<td>Ordinal</td>
</tr>
<tr>
<td>8</td>
<td>Dual incomes</td>
<td>3</td>
<td>Categorical</td>
</tr>
<tr>
<td>9</td>
<td>Number in household</td>
<td>9</td>
<td>Ordinal</td>
</tr>
<tr>
<td>10</td>
<td>Number of children</td>
<td>9</td>
<td>Ordinal</td>
</tr>
<tr>
<td>11</td>
<td>Householder status</td>
<td>3</td>
<td>Categorical</td>
</tr>
<tr>
<td>12</td>
<td>Type of home</td>
<td>5</td>
<td>Categorical</td>
</tr>
<tr>
<td>13</td>
<td>Ethnic classification</td>
<td>8</td>
<td>Categorical</td>
</tr>
<tr>
<td>14</td>
<td>Language in home</td>
<td>3</td>
<td>Categorical</td>
</tr>
</tbody>
</table>
Toy Example

An itemset is a pattern defined by

\((A_{i1} = a_{j1}) \land (A_{i2} = a_{j2}) \land \cdots \land (A_{ik} = a_{jk})\)

The frequency (or support) of an itemset \(X\) is simply \(P(X)\).

Example: in the "Play Tennis" data

\(P(\text{Humidity} = \text{Normal} \land \text{Play} = \text{Yes} \land \text{Windy} = \text{False}) = \frac{4}{14}\)

The accuracy (or confidence) of an association rule \(\text{if } Y=y \Rightarrow Z=z\) is

\(P(Z = z \mid Y = y)\).

Example

\(P(\text{Windy} = \text{False} \Rightarrow \text{Play} = \text{Yes} \mid \text{Humidity} = \text{Normal}) = \frac{4}{7}\)

Generating rules from itemsets

An itemset of size \(k\) can give rise to \(2^k\) rules.

Example. Itemset \(\text{Windy=False, Play=Yes, Humidity=Normal}\) gives rise to 7 rules including:

- \(\text{IF Windy=False and Humidity=Normal THEN Play=Yes (4/4)}\)
- \(\text{IF Play=Yes THEN Humidity=Normal and Windy=False (4/9)}\)
- \(\text{IF True THEN Windy=False and Play=Yes and Humidity=Normal (4/14)}\)

Select association rules that have accuracy greater than some threshold.

Finding Frequent Itemsets

Task: find all itemsets with frequency \(s\).

Key observation: a set \(X\) of variables can be frequent only if all subsets of variables are frequent (monotonicity property), i.e.

\(P(A,B) = P(A) \land P(A,B) = P(B)\).

So find frequent singleton sets, then sets of size 2, and so on...

An efficient algorithm using this idea for finding frequent itemsets is the APRIORI algorithm (Agrawal and Srikant (1994), Mannila et al (1994)).
Itemsets, Coverage, etc

• Call each column an attribute $A_1, A_2, \ldots A_m$

• An item set is a set of attribute value pairs

$$\{A_{i_1} = a_{j_1}, A_{i_2} = a_{j_2}, \ldots, A_{i_k} = a_{j_k}\}$$

• Example: In the Play Tennis data

Humidity = Normal $\land$ Play = Yes $\land$ Windy = False

• The support of an item set is its frequency in the data set

• Example:

**support** ($\text{Humidity = Normal } \land \text{ Play = Yes } \land \text{ Windy = False}$) = 4

• The confidence of an association rule if $Y=y$ then $Z=z$ is

$$P(Z = z | Y = y)$$

• Example:

$$P(\text{Windy = False } \land \text{ Play = Yes } | \text{ Humidity = Normal}) = 4/7$$
Item sets to rules

- First: We will find frequent item sets
- Then: We convert them to rules
- An itemset of size \( k \) can give rise to \( 2^k - 1 \) rules
- Example: itemset
  
  \[ \text{Windy}=\text{False}, \ \text{Play}=\text{Yes}, \ \text{Humidity}=\text{Normal} \]

- Results in 7 rules including:

  IF Windy=\text{False} and Humidity=\text{Normal} THEN Play=\text{Yes} \quad (4/4)
  IF Play=\text{Yes} THEN Humidity=\text{Normal} and Windy=\text{False} \quad (4/9)
  IF True THEN Windy=\text{False} and Play=\text{Yes} and Humidity=\text{Normal} \quad (4/14)

- We keep rules only whose confidence is greater than a threshold
Finding Frequent Itemsets

• Task: Find all item sets with support

• Insight: A large set can be no more frequent than its subsets, e.g.,

  \[ \text{support}(\text{Wind} = \text{False}) \geq \text{support}(\text{Wind} = \text{False}, \text{Outlook} = \text{Sunny}) \]

• So search through itemsets in order of number of items

• An efficient algorithm for this is \textsc{Apriori} (Agarwal and Srikant, 1994; Mannila et al, 1994)
APRIORI Algorithm

(for binary variables)

\[ i = 1 \]

\[ C_i = \{\{A\}|A \text{ is a variable}\} \]

while \( C_i \) is not empty

database pass:

for each set in \( C_i \) test if it is frequent

let \( L_i \) be collection of frequent sets from \( C_i \)

candidate formation:

let \( C_{i+1} \) be those sets of size \( i + 1 \)

all of whose subsets are frequent

end while
Single database pass is linear in $|C_i|n$, make a pass for each $i$ until $C_i$ is empty

Candidate formation

- Find all pairs of sets $\{U, V\}$ from $L_i$ such that $U \cup V$ has size $i + 1$ and test if this union is really a potential candidate. $O(|L_i|^3)$

Example: 5 three-item sets
(ABC), (ABD), (ACD), (ACE), (BCD)
Candidate four-item sets
(ABCD) ok
(ACDE) not ok because (CDE) is not present above
Comments

- Some association rules will be trivial, some interesting. Need to sort through them
  - Example: pregnant => female (confidence: 1)
- Also can miss “interesting but rare” rules
  - Example: vodka --> caviar (low support)
- Really this is a type of exploratory data analysis
- For rule A --> B, can be useful to compare \( P(B|A) \) to \( P(B) \)
- \textsc{apriori} can be generalised to structures like subsequences and subtrees